Section 3

Iterative Methods in Matrix Algebra

Vector norm

Definition

A **vector norm** *on* $\mathbb{R}^n$*, denoted by* $\|\cdot\|$*, is a mapping from* $\mathbb{R}^n$ *to* $\mathbb{R}$ *such that*

- $\|x\| \geq 0$ *for all* $x \in \mathbb{R}^n$*,*
- $\|x\| = 0$ *if and only if* $x = 0$*,*
- $\|\alpha x\| = |\alpha|\|x\|$ *for all* $\alpha \in \mathbb{R}$ *and* $x \in \mathbb{R}^n$*,*
- $\|x + y\| \leq \|x\| + \|y\|$ *for all* $x, y \in \mathbb{R}$*.*

Vector norm

Definition (l_p norms)

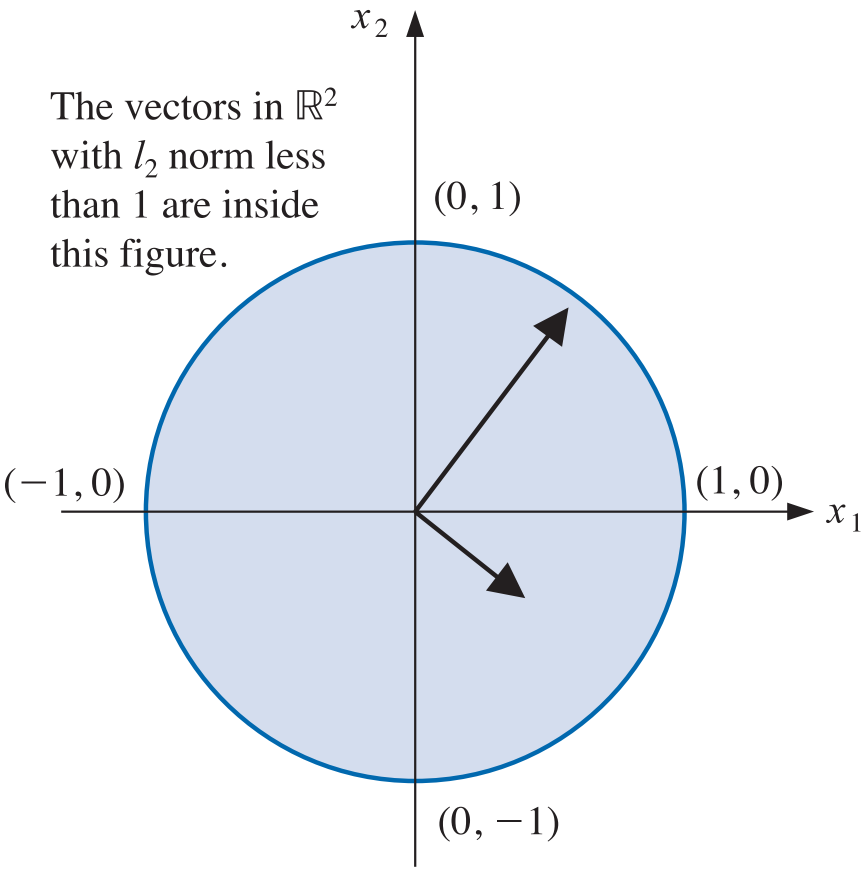
The l_p (sometimes L_p or ℓ_p) norm of a vector is defined by

$$
\begin{aligned}
1 \le p < \infty : &\qquad \|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p} \\
p = \infty : &\qquad \|x\|_\infty = \max_{1\le i\le n} |x_i|
\end{aligned}
$$

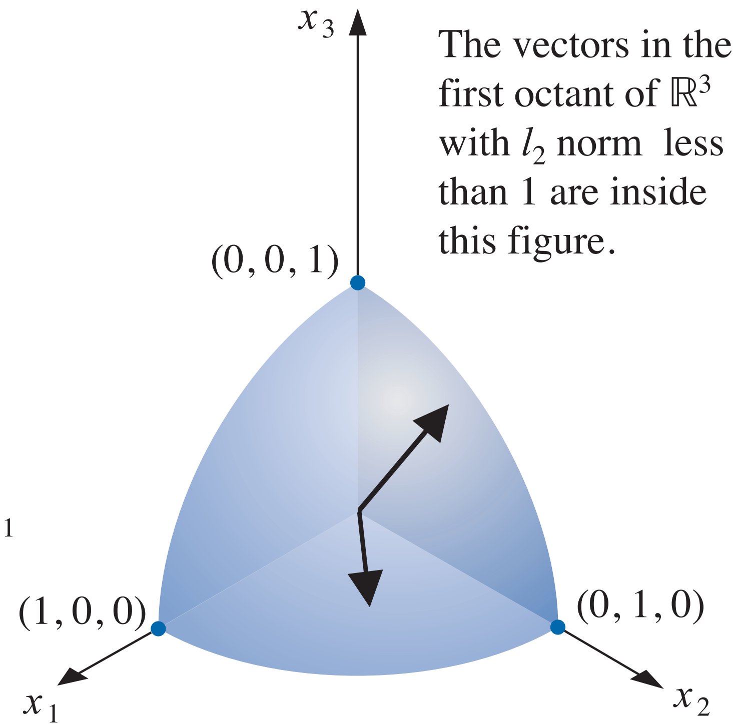
In particular, the l_2 norm is also called the **Euclidean norm**.

Note that when $0 \le p < 1$, $\|\cdot\|_p$ is not norm, strictly speaking, but have some usages in specific applications.

l_2 norm

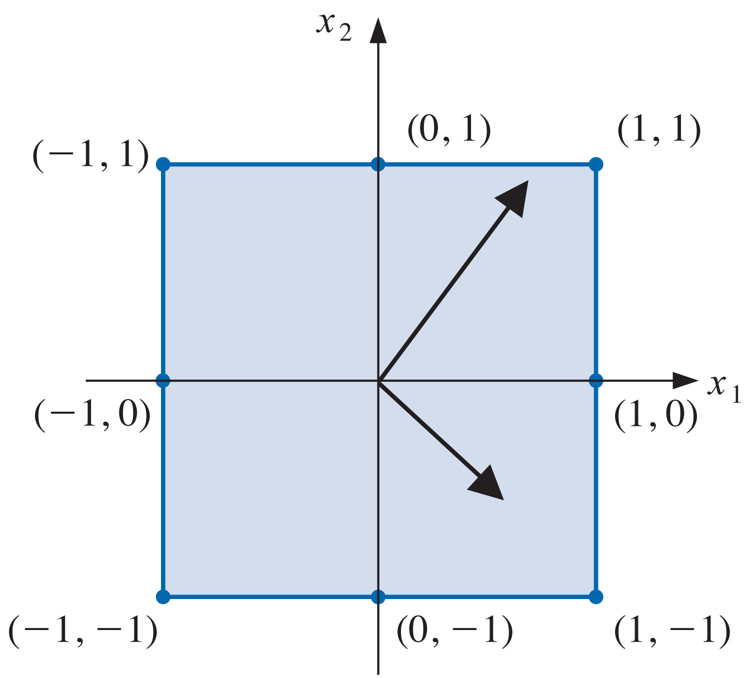


x_2
The vectors in $\mathbb{R}^2$ with l_2 norm less than 1 are inside this figure.
(0, 1)
(−1, 0)
(1, 0)
x_1
(0, −1)

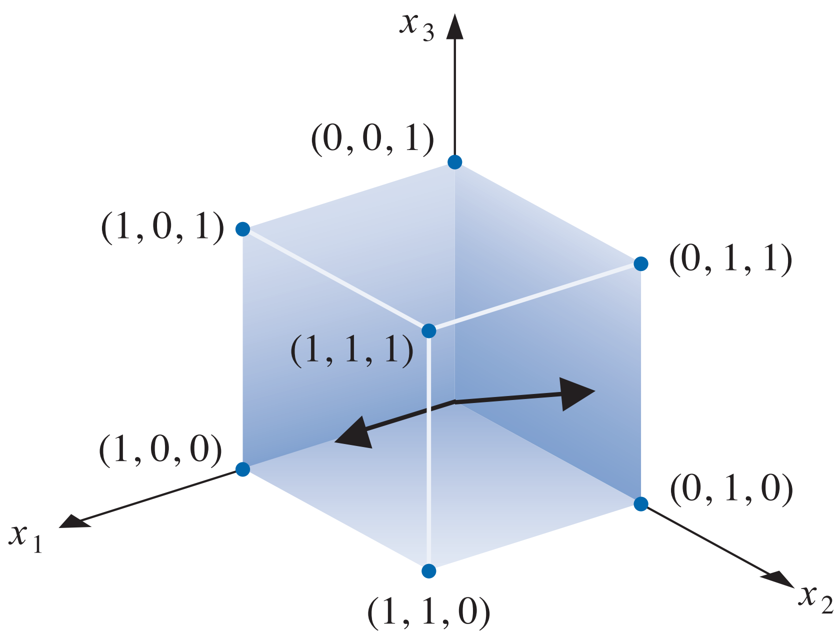


x_3
The vectors in the first octant of $\mathbb{R}^3$ with l_2 norm less than 1 are inside this figure.
(0, 0, 1)
(1, 0, 0)
(0, 1, 0)
x_1
x_2

l_∞ norm



The vectors in $\mathbb{R}^2$ with l_∞ norm less than 1 are inside this figure.



The vectors in the first octant of $\mathbb{R}^3$ with l_∞ norm less than 1 are inside this figure.

Vector norms

Example

Compute the l_2 and l_∞ norms of vector $x = (1, -1, 2) \in \mathbb{R}^3$.

Solution:

$$\|x\|_2 = \sqrt{|1|^2 + |-1|^2 + |2|^2} = \sqrt{6}$$

$$\|x\|_\infty = \max_{1 \le i \le 3} |x_i| = \max\{|1|, |-1|, |2|\} = 2$$

Theorem (Cauchy-Schwarz inequality)

For any vectors $x = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$ *and* $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$*, there is*

$$|x^\top y| = \left|\sum_{i=1}^n x_i y_i\right| \le \left(\sum_{i=1}^n |x_i|^2\right)^{1/2} \left(\sum_{i=1}^n |y_i|^2\right)^{1/2} = \|x\|_2 \|y\|_2$$

Proof.

It is obviously true for $x = 0$ or $y = 0$. If $x, y \ne 0$, then for any $\lambda \in \mathbb{R}$, there is

$$0 \le \|x - \lambda y\|_2^2 = \|x\|_2^2 - 2\lambda x^\top y + \lambda^2 \|y\|_2^2$$

and the equality holds when $\lambda = \|x\|_2 / \|y\|_2$. $\square$

Distance between vectors

Definition (Distance between two vectors)

The l_p **distance** $(1 \leq p \leq \infty)$ *between two vectors* $x, y \in \mathbb{R}^n$ *is defined by* $\|x - y\|_p$.

Definition (Convergence of a sequence of vectors)

A sequence $\{x^{(k)}\}$ *is said to* **converge with respect to the** l_p **norm** *if for any given* $\epsilon > 0$, *there exists an integer* $N(\epsilon)$ *such that*

$$\|x^{(k)} - x\| < \epsilon, \quad \text{for all } k \geq N(\epsilon)$$

Convergence of a sequence of vectors

Theorem

A sequence of vectors $\{x^{(k)}\}$ converges to x if and only if $x_i^{(k)} \to x_i$ for every $i = 1, 2, \ldots, n$.

Theorem

For any vector $x \in \mathbb{R}^n$, there is

$$\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty$$

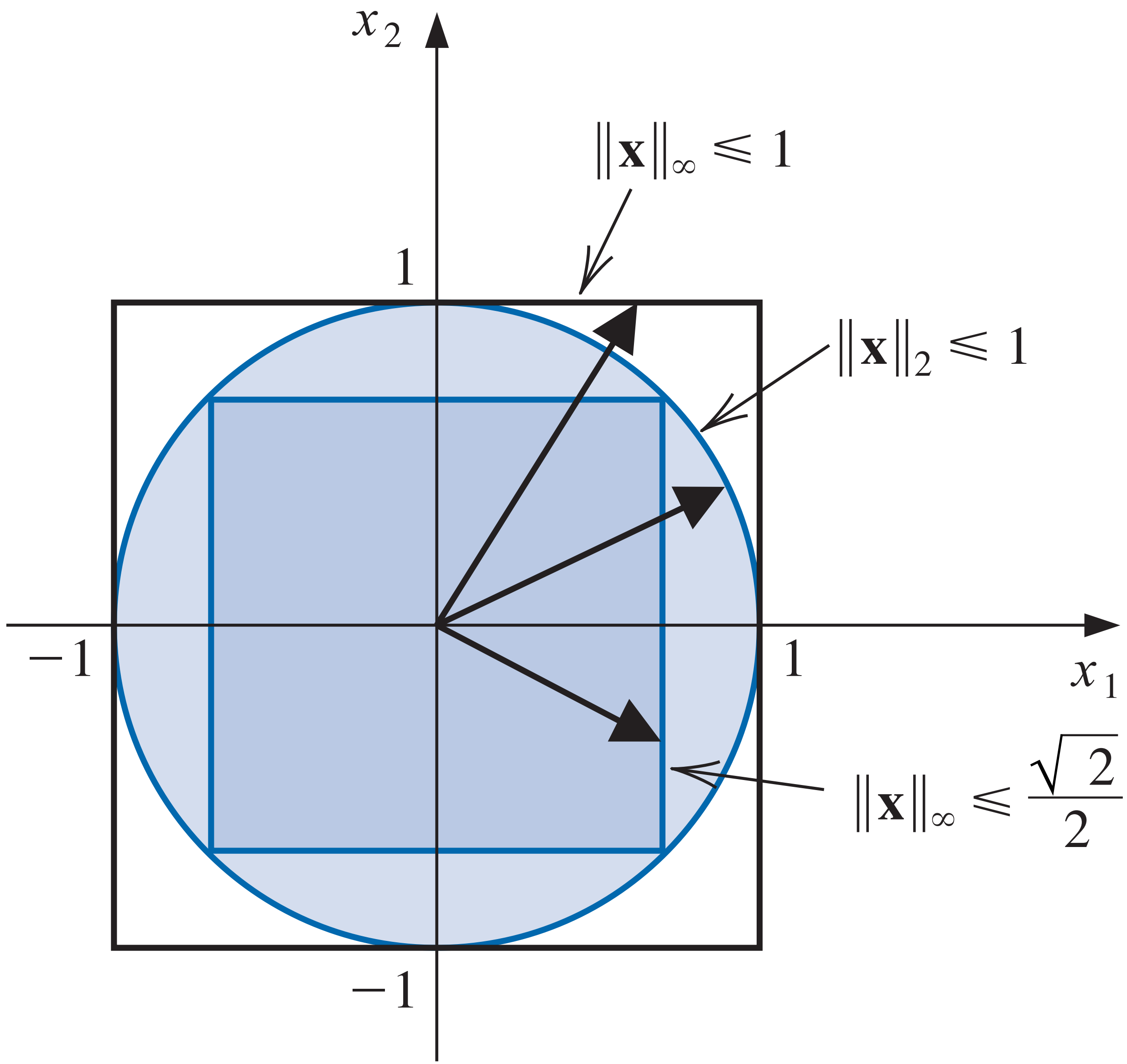
Proof.

$$\|x\|_\infty = \max_i |x_i| = \sqrt{\max_i |x_i|^2} \le \sqrt{|x_1|^2 + \cdots + |x_n|^2} = \|x\|_2$$

$$\begin{aligned}\|x\|_2 &= \sqrt{|x_1|^2 + \cdots + |x_n|^2} \le \sqrt{n \max_i |x_i|^2} \\ &= \sqrt{n}\sqrt{\max_i |x_i|^2} = \sqrt{n} \max_i |x_i| = \sqrt{n}\|x\|_\infty\end{aligned}$$

□

Compare l_2 and l_∞ norms in $\mathbb{R}^2$



Matrix norm

Definition

A **matrix norm** *on the set of $n \times n$ matrices is a real-valued function, denoted by $\|\cdot\|$, that satisfies the follows for all $A, B \in \mathbb{R}^{n\times n}$ and $\alpha \in \mathbb{R}$:*

- $\|A\| \geq 0$
- $\|A\| = 0$ *if and only if* $A = 0$ *the zero matrix,*
- $\|\alpha A\| = |\alpha|\|A\|$
- $\|A + B\| \leq \|A\| + \|B\|$
- $\|AB\| \leq \|A\|\|B\|$

Distance between matrices

Definition

Suppose $\|\cdot\|$ is a norm defined on $\mathbb{R}^{n\times n}$. Then the **distance between two** $n \times n$ **matrices** A **and** B *with respect to $\|\cdot\|$ is $\|A - B\|$ (check that it's a distance)*

Matrix norm can be induced by vector norms, and hence there are many choices. Here we focus on those induced by l_2 and l_∞ vector norms.

Matrix norm

Definition

If $\|\cdot\|$ is a vector norm on $\mathbb{R}^n$, then the norm defined below

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$

is called the **matrix norm induced by vector norm** $\|\cdot\|$.

Matrix norm

Remark

- *Induced norms are also called natural norms of matrices.*
- *Unless otherwise specified, by matrix norms most books/papers refer to induced norms.*
- *The induced norm can be written equivalently as*

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

- *It can be easily extended to case $A \in \mathbb{R}^{m \times n}$.*

Matrix norm

Corollary

For any vector $x \in \mathbb{R}^n$, there is $\|Ax\| \leq \|A\|\|x\|$.

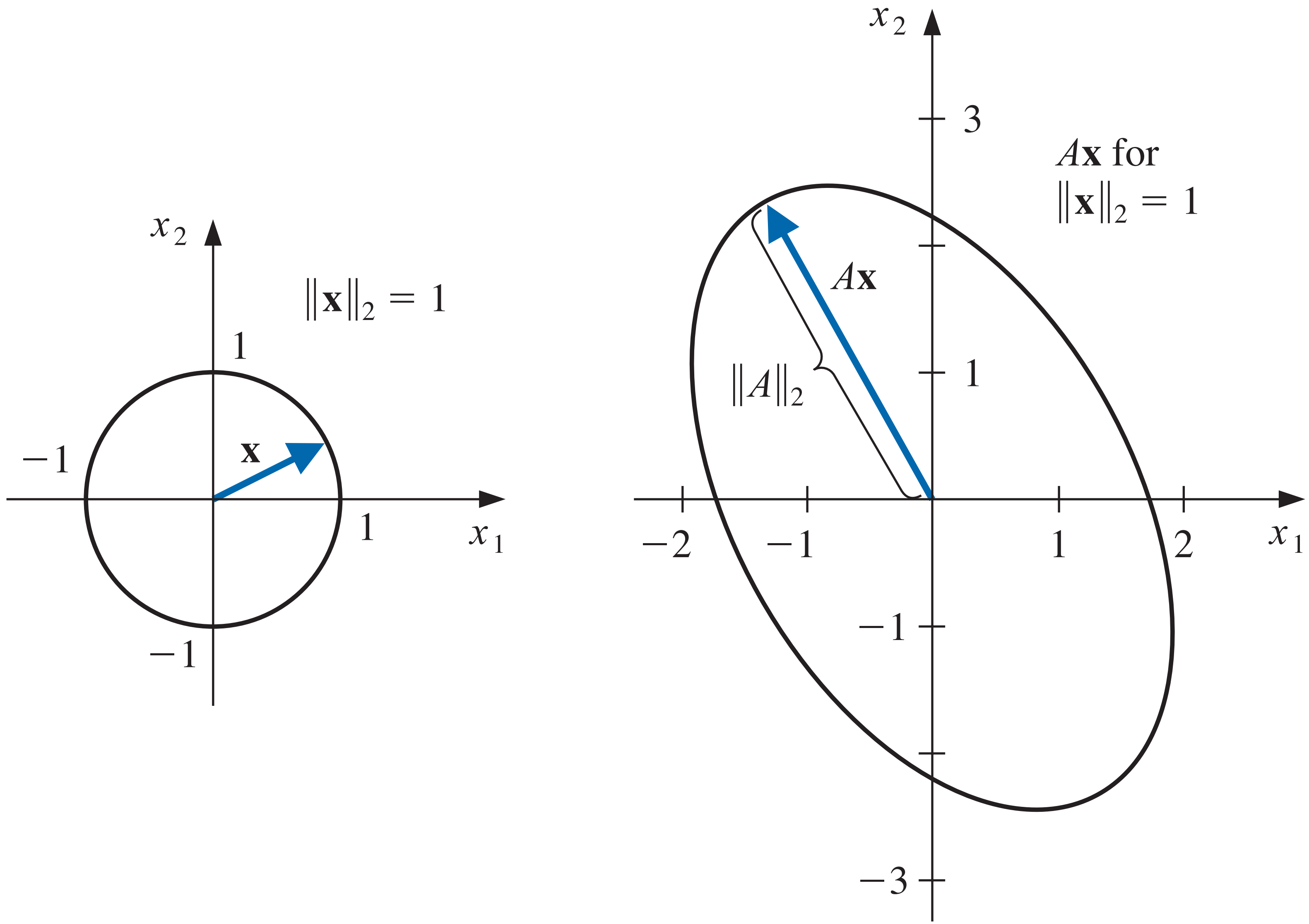
Proof.

It is obvious for $x = 0$. If $x \neq 0$, then

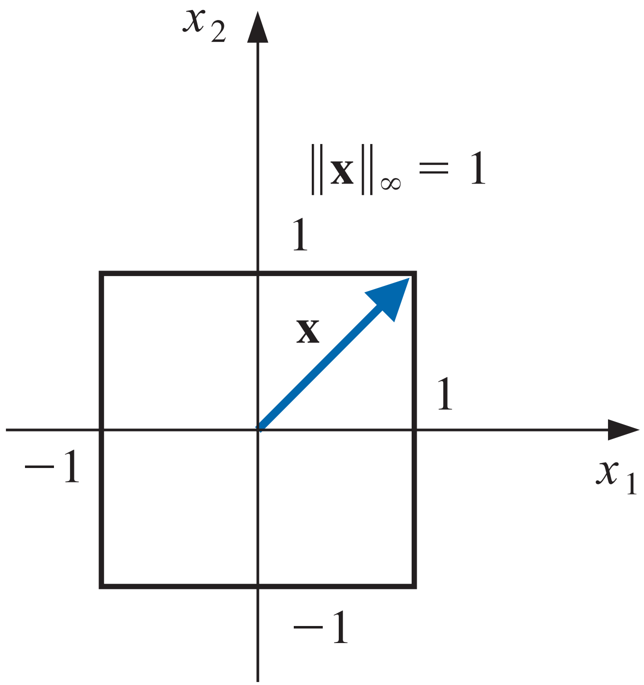
$$\frac{\|Ax\|}{\|x\|} \leq \max_{x' \neq 0} \frac{\|Ax'\|}{\|x'\|} = \|A\|$$

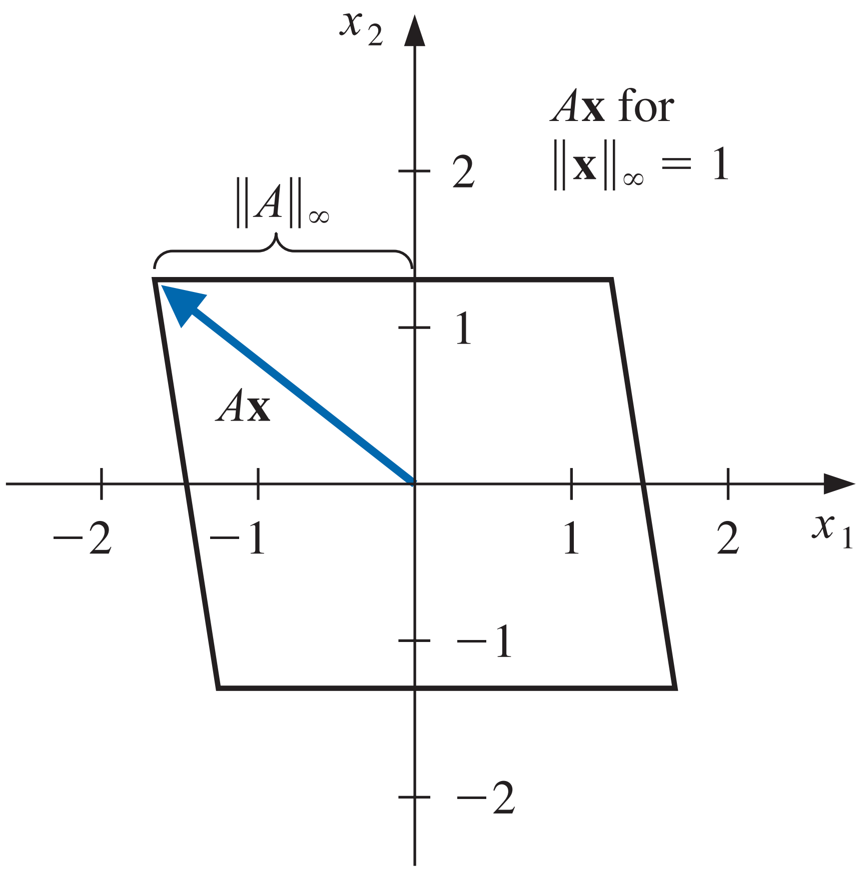
$\square$

Induced l_2 matrix norm



Induced l_∞ matrix norm





Matrix norm

Theorem

Suppose $A = [a_{ij}] \in \mathbb{R}^{n\times n}$*, then* $\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^n |a_{ij}|$.

Matrix norm

Proof.

For any x with $\|x\|_\infty = 1$, i.e., $\max_i |x_i| = 1$, there is

$$\begin{aligned}
\|Ax\|_\infty &= \max \left\{ \left| \sum_j a_{1j} x_j \right|, \dots, \left| \sum_j a_{nj} x_j \right| \right\} \\
&\le \max \left\{ \sum_j |a_{1j}| |x_j|, \dots, \sum_j |a_{nj}| |x_j| \right\} \\
&\le \max \left\{ \sum_j |a_{1j}|, \dots, \sum_j |a_{nj}| \right\} = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|
\end{aligned}$$

Suppose i' is such that $\sum_{j=1}^n |a_{i'j}| = \max_{1 \le i \le n} \sum_{j=1}^n |a_{ij}|$, then by choosing $\hat{x}$ such that $\hat{x}_j = 1$ if $a_{i'j} \ge 0$ and -1 otherwise, we have $\sum_{j=1}^n a_{i'j} \hat{x}_j = \sum_{j=1}^n |a_{i'j}|$. So $\|A\hat{x}\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^n |a_{ij}|$. Note that $\|\hat{x}\|_\infty = 1$. Therefore $\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^n |a_{ij}|$. $\square$

Eigenvalues and eigenvectors of square matrices

Definition

The **characteristic polynomial** *of a square matrix* $A \in \mathbb{R}^{n\times n}$ *is defined by*

$$p(\lambda) = \det(A - \lambda I)$$

We call λ *an* **eigenvalue** *of* A *if* λ *is a root of* p*, i.e.,* $\det(A - \lambda I) = 0$*. Moreover, any nonzero solution* $x \in \mathbb{R}^n$ *of* $(A - \lambda I)x = 0$ *is called an* **eigenvector** *of* A *corresponding to the eigenvalue* λ*.*

Eigenvalues and eigenvectors of square matrices

Remark

- *$p(\lambda)$ is a polynomial of degree n, and hence has n roots.*
- *x is an eigenvector of A corresponding to eigenvalue λ iff $(A - \lambda I)x = 0$, i.e., $Ax = \lambda x$. This also means A applied to x is stretching x by λ.*
- *If x is an eigenvector of A corresponding to λ, so is αx for any $\alpha \neq 0$:*
$$A(\alpha x) = \alpha Ax = \alpha\lambda x = \lambda(\alpha x)$$

Eigenvalues and eigenvectors of square matrices

Definition

Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $A \in \mathbb{R}^{n\times n}$, then the **spectral radius** *$\rho(A)$ is defined by $\rho(A) = \max_i |\lambda_i|$ where $|\cdot|$ is the absolute value (aka magnitude) of complex numbers.*

Eigenvalues and eigenvectors of square matrices

Some properties

Theorem

For a matrix $A \in \mathbb{R}^{n\times n}$, there are

- $\|A\|_2 = \sqrt{\rho(A^\top A)}$
- $\rho(A) \le \|A\|$ *for any norm* $\|\cdot\|$ *of* A

Proof.

- We later will show that both sides $= \sigma_1^2$, where σ_1 is the largest singular value of A.
- Let $\lambda := \rho(A)$ be the eigenvalue with largest magnitude. Then there exists eigenvector x such that

$$(\|A\| \ge) \frac{\|Ax\|}{\|x\|} = \frac{\|\lambda x\|}{\|x\|} = \frac{|\lambda|\|x\|}{\|x\|} = |\lambda|$$

$\square$

Convergent matrix

Definition

A matrix $A \in \mathbb{R}^{n\times n}$ is said to be **convergent** *if*

$$\lim_{k\to\infty} A^k = 0$$

Theorem

The following statements are equivalent:

1. *A is convergent.*
2. $\lim_{k\to\infty} \|A^k\| = 0$ *for any norm* $\|\cdot\|$.
3. $\rho(A) < 1$.
4. $\lim_{k\to\infty} A^k x = 0$ *for any* $x \in \mathbb{R}$.

Jacobi iterative method

To solve x from $Ax = b$ where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, the **Jacobi iterative method** is

- Initialize $x^{(0)} \in \mathbb{R}^n$. Set $D = \mathrm{diag}(A)$, $R = A - D$.
- Repeat the following for $k = 0, 1, \dots$ until convergence:

$$x^{(k+1)} = D^{-1}(b - Rx^{(k)})$$

Remark

- *Needs nonzero diagonal entries, i.e., $a_{ii} \neq 0$ for all i.*
- *Usually faster convergence with larger $|a_{ii}|$.*
- *Stopping criterion can be $\frac{\|x^{(k)}-x^{(k-1)}\|}{\|x^{(k)}\|} \le \epsilon$ for some prescribed $\epsilon > 0$.*

Gauss-Seidel iterative method

To solve x from $Ax = b$ where $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$, the **Gauss-Seidel iterative method** is

- Initialize $x^{(0)} \in \mathbb{R}^n$. Set L to the lower triangular part (including diagonal) of A and $U = A - L$.
- Repeat the following for $k = 0, 1, \ldots$ until convergence:

$$x^{(k+1)} = L^{-1}(b - Ux^{(k)})$$

Remark

- *Inverse of L requires forward substitution.*
- *Again needs nonzero diagonal entries, i.e., $a_{ii} \neq 0$ for all i.*
- *Stopping criterion can be $\frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^{(k)}\|} \leq \epsilon$ for some prescribed $\epsilon > 0$.*
- *Faster than Jacobi iterative method most of times.*

General iterative methods

Lemma ($\rho(T) < 1 \Rightarrow I - T$ invertible)

If $\rho(T) < 1$, then $(I - T)^{-1}$ exists and

$$(I - T)^{-1} = I + T + T^2 + \cdots = \sum_{j=0}^{\infty} T^j$$

General iterative methods

Proof.

We first show that $I - T$ is invertible, i.e., $(I - T)x = 0$ has unique solution $x = 0$. If not, then $\exists x \neq 0$ such that $(I - T)x = 0$, i.e., $Tx = x$, or x is an e.v. corresponding to e.w. 1, contradiction to $\rho(T) < 1$.

Define $S_m = I + T + \cdots + T^m$. Then $(I - T)S_m = I - T^{m+1}$. Note $\rho(T) < 1$ implies $\lim_{m\to\infty} T^m = 0$, and hence

$$(I - T)\lim_{m\to\infty} S_m = \lim_{m\to\infty}(I - T)S_m = \lim_{m\to\infty}(I - T^{m+1}) = I$$

That is, $\sum_{m=0}^{\infty} T^m = \lim_{m\to\infty} S_m = (I - T)^{-1}$. $\square$

General iterative methods

General iterative method has form $x^{(k)} = Tx^{(k-1)} + c$ for $k = 1, 2, \dots$.

Example (Jacobi and GS are iterative methods)

- Jacobi iterative method:

$$x^{(k)} = D^{-1}(b - Rx^{(k-1)}) = -(D^{-1}R)x^{(k-1)} + D^{-1}b$$

  So $T = -D^{-1}R$ and $c = D^{-1}b$.

- Gauss-Seidel iterative method:

$$x^{(k)} = L^{-1}(b - Ux^{(k-1)}) = -(L^{-1}U)x^{(k-1)} + L^{-1}b$$

  So $T = -L^{-1}U$ and $c = L^{-1}b$.

General iterative methods

Theorem (Sufficient and necessary condition of convergence)

For any initial $x^{(0)}$, the sequence $\{x^{(k)}\}_k$ defined by

$$x^{(k)} = Tx^{(k-1)} + c$$

converges to the unique solution of $x = Tx + c$ iff $\rho(T) < 1$.

Proof.

($\Leftarrow$) Suppose $\rho(T) < 1$. Then

$$\begin{aligned} x^{(k)} &= Tx^{(k-1)} + c = T(Tx^{(k-2)} + c) + c = T^2 x^{(k-2)} + (I + T)c \\ &= \cdots = T^k x^{(0)} + (I + T + \cdots + T^k)c \end{aligned}$$

Note $\rho(T) < 1 \Rightarrow T^k \to 0$ and $(I + T + \cdots + T^k) \to (I - T)^{-1}$, so $x^{(k)} \to (I - T)^{-1}c$, the unique solution of $x = Tx + c$. $\square$

General iterative methods

Proof.

($\Rightarrow$) Let x^* be the unique solution of $x = Tx + c$. Then for any $z \in \mathbb{R}^n$, we set initial $x^{(0)} = x^* - z$. Then

$$\begin{aligned} x^* - x^{(k)} &= (Tx^* + c) - (Tx^{(k-1)} + c) = T(x^* - x^{(k-1)}) \\ &= \cdots = T^k(x^* - x^{(0)}) = T^k z \to 0 \end{aligned}$$

This implies $\rho(T) < 1$. $\square$

General iterative methods

Corollary (Linear convergence rate)

If $\|T\| < 1$ for any matrix norm $\|\cdot\|$, and c is given, then $\{x^{(k)}\}$ generated by $x^{(k)} = Tx^{(k-1)} + c$ converges to the unique solution x^* of $x = Tx + c$. Moreover

1. $\|x^* - x^{(k)}\| \le \|T\|^k \|x^* - x^{(0)}\|$.
2. $\|x^* - x^{(k)}\| \le \frac{\|T\|^k}{1-\|T\|} \|x^{(1)} - x^{(0)}\|$.

Proof.

1. Note $\rho(T) \le \|T\| < 1$. Follow ($\Rightarrow$) part of the theorem above.
2. Note that $\|x^* - x^{(1)}\| \le \|T\|\|x^* - x^{(0)}\|$ and hence $\|x^{(1)} - x^{(0)}\| \ge \|x^* - x^{(0)}\| - \|x^* - x^{(1)}\| \ge (1 - \|T\|)\|x^* - x^{(0)}\|$.

$\square$

General iterative methods

Theorem (Jacobi and GS are convergent)

If A is strictly diagonally dominant, then from any initial $x^{(0)}$ both Jacobi and Gauss-Seidel iterative methods generate sequences that converge to the unique solution of $Ax = b$.

Proof.

For Jacobi, we can show $\rho(D^{-1}R) < 1$: if not, then exists ew λ such that $|\lambda| = \rho(D^{-1}R) \geq 1$, and ev $x \neq 0$ such that $D^{-1}Rx = \lambda x$, i.e., $(R + \lambda D)x = 0$ or $R + \lambda D$ invertible, contradiction to $A = D + R$ strictly diagonally dominant given $|\lambda| \geq 1$. Similar for GS. $\square$

Relaxation techniques

The theory of general iterative methods suggest using a matrix T with smaller spectrum $\rho(T)$. To this end, we can use the relaxation technique to modify the iterative scheme.

- Original Gauss-Seidel iterative method:

$$x^{(k)} = -(L^{-1}U)x^{(k-1)} + L^{-1}b$$

- **Successive Over-Relaxation**[1] (SOR) for Gauss-Seidel iterative method ($\omega > 1$):

$$x^{(k)} = (D - \omega L)^{-1}[(1-\omega)D + \omega U]x^{(k-1)} + \omega(D - \omega L)^{-1}b$$

  where $D, -L, -U$ are the diagonal, strict lower, and strict upper triangular parts of A, respectively.

---

[1] $Ax = b \Leftrightarrow \omega(-L + D - U)x = \omega b \Leftrightarrow (D - \omega L)x = ((1-\omega)D + \omega U)x + \omega b.$

Relaxation techniques

Example

Compare Gauss-Seidel and SOR with $\omega = 1.25$, both using $x^{(0)} = (1, 1, 1)^\top$ as initial, to solve the system:

$$
\begin{aligned}
4x_1 + 3x_2 \qquad &= 24 \\
3x_1 + 4x_2 - x_3 &= 30 \\
-x_2 + 4x_3 &= -24
\end{aligned}
$$

Relaxation techniques

Solution: Compare with true solution $(3, 4, -5)^{\top}$, we get:

Gauss-Seidel:

k	0	1	2	3	4	5	6	7
$x_1^{(2)}$	1	5.250000	3.1406250	3.0878906	3.0549316	3.0343323	3.0214577	3.0134110
$x_2^{(2)}$	1	3.812500	3.8828125	3.9667578	3.9542236	3.9713898	3.9821186	3.9888241
$x_3^{(2)}$	1	-5.046875	-5.0292969	-5.0183105	-5.0114441	-5.0071526	-5.0044703	-5.0027940

Successive Over-Relaxation:

k	0	1	2	3	4	5	6	7
$x_1^{(k)}$	1	6.312500	2.6223145	3.1333027	2.9570512	3.0037211	2.9963276	3.0000498
$x_2^{(k)}$	1	3.5195313	3.9585266	4.0102646	4.0074838	4.0029250	4.0009262	4.0002586
$x_3^{(k)}$	1	-6.6501465	-4.6004238	-5.0966863	-4.9734897	-5.0057135	-4.9982822	-5.0003486

The 5th iteration of SOR is better than 7th of GS.

Relaxation techniques

Theorem (Kahan's theorem)

If all diagonal entries of A are nonzero, then $\rho(T_\omega) \ge |\omega - 1|$, where $T_\omega = (D - \omega L)^{-1}[(1-\omega)D + \omega U]$.

Proof.

Let $\lambda_1, \dots, \lambda_n$ be the ew of T_ω, then

$$\prod_{i=1}^{n} \lambda_i = \det(T_\omega) = \det(D)^{-1} \det((1-\omega)D) = (1-\omega)^n$$

since $D - \omega L$ and $(1-\omega)D + \omega U$ are lower/upper triangular matrices. Hence $\rho(T_\omega)^n \ge \prod_{i=1}^{n} |\lambda_i| = |1-\omega|^n$. $\square$

This result says that SOR can converge only if $|\omega - 1| < 1$.

Relaxation techniques

Theorem (Ostrowski-Reich theorem)

If A is positive definite and $|\omega - 1| < 1$, then the SOR converges starting from any initial $x^{(0)}$.

Theorem

If A is positive definite and tridiagonal, then $\rho(T_g) = [\rho(T_j)]^2 < 1$, where T_g and T_j are the T matrices of GS and Jacobi methods respectively, and the optimal ω for SOR is

$$\omega = \frac{2}{1 + \sqrt{1 - (\rho(T_j))^2}}$$

With this choice of ω, the spectrum $\rho(T_\omega) = \omega - 1$.

Iterative refinement

Definition (Residual)

Let $\tilde{x}$ be an approximation to the solution x of linear system $Ax = b$. Then $r = b - A\tilde{x}$ is called the **residual** *of approximation $\tilde{x}$.*

Remark

It seems intuitive that a small residual r implies a close approximation $\tilde{x}$ to x. However, it is not always true.

Iterative refinement

Example (small residual $\nRightarrow$ small approximation error)

The linear system $Ax = b$ is given by

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}$$

has a unique solution $x = (1, 1)^\top$. Determine the residual vector r of a poor approximation $\tilde{x} = (3, -0.0001)^\top$.

Solution: The residual is

$$r = b - A\tilde{x} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -0.0001 \end{bmatrix} = \begin{bmatrix} 0.0002 \\ 0 \end{bmatrix}$$

So $\|r\|_\infty = 0.0002$ is small but $\|\tilde{x} - x\|_\infty = 2$ is large.

Iterative refinement

Theorem (Relation between residual and error)

Suppose A is nonsingular, and $\tilde{x}$ is an approximation to the solution x of $Ax = b$, and $r = b - A\tilde{x}$ is the residual vector of $\tilde{x}$, then for any norm, there is

$$\|x - \tilde{x}\| \le \|r\| \cdot \|A^{-1}\|$$

Moreover, if $x \neq 0$ and $b \neq 0$, then there is

$$\frac{\|x - \tilde{x}\|}{\|x\|} \le \|A\| \cdot \|A^{-1}\| \cdot \frac{\|r\|}{\|b\|}$$

If $\|A\|\|A^{-1}\|$ is large, then small $\|r\|$ does not guarantee small $\|x - \tilde{x}\|$.

Iterative refinement

Proof.

Since x is a solution, we have $Ax = b$, we have $r = b - A\tilde{x} = Ax - A\tilde{x} = A(x - \tilde{x})$. Since A is nonsingular, we have $x - \tilde{x} = A^{-1}r$, and hence

$$\|x - \tilde{x}\| = \|A^{-1}r\| \leq \|r\| \cdot \|A^{-1}\|$$

If $x \neq 0$ and $b \neq 0$, from $\|b\| = \|Ax\| \leq \|A\| \cdot \|x\|$ we have $1/\|x\| \leq \|A\|/\|b\|$. Multiplying this to the inequality above, we get

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \cdot \frac{\|r\|}{\|b\|}$$

$\square$

Iterative refinement

The number $\|A\| \cdot \|A^{-1}\|$ provide an indication between the error of approximation $\|x - \tilde{x}\|$ and size of residual r. So the larger $\|A\| \cdot \|A^{-1}\|$ is, the less power we have to control error using residual.

Definition (Condition number)

The **condition number** *of a nonsingular matrix A relative to a norm $\|\cdot\|_p$ is*

$$K_p(A) = \|A\|_p \cdot \|A^{-1}\|_p$$

The subscript p is often omitted if it's clear from context or it's not important.

Condition number

Remark

- *The condition number $K(A) \geq 1$:*

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \cdot \|A^{-1}\| = K(A)$$

- *A matrix A is called* **well-conditioned** *if $K(A)$ is close to 1.*
- *A matrix A is called* **ill-conditioned** *if $K(A) \gg 1$.*

Condition number

Example (Condition number)

Determine the condition number of matrix

$$A = \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}$$

Condition number

Solution: Let's use l_∞ norm. Then

$$\|A\|_\infty = \max\{|1| + |2|, |1.0001| + |2|\} = 3.0001$$

Furthermore, there is

$$A^{-1} = \begin{bmatrix} -10000 & 10000 \\ 5000.5 & -5000 \end{bmatrix}$$

and hence $\|A^{-1}\|_\infty = 20000$. Therefore

$$K(A) = \|A\| \cdot \|A^{-1}\| = 3.0001 \times 20000 = 60002$$

Iterative refinement

Suppose $\tilde{x}$ is our current approximation to x. Let $\tilde{y} = x - \tilde{x}$, then $A\tilde{y} = A(x - \tilde{x}) = Ax - A\tilde{x} = b - A\tilde{x} = r$. If we can solve for $\tilde{y}$ here, we would get a new approximation $\tilde{x} + \tilde{y}$, expectedly to approximate x better.

This procedure is called **iterative refinement**.

Iterative refinement

Given A and b, Iterative Refinement first applies Gauss eliminations to $Ax = b$ and obtains approximation x.

Then, for each iteration $k = 1, 2, \dots, N$, do the following:

- Compute residual $r = b - Ax$;
- Solve y from $Ay = r$ using the same Gauss elimination steps.
- Set $x \leftarrow x + y$

The actual Iterative Refinement algorithm can also find approximation of condition number $K_\infty(A)$ (See textbook).

Perturbed linear system

In reality, A and b may be perturbed by noise or rounding errors δA and δb. Therefore, we are actually solving

$$(A + \delta A)x = b + \delta b$$

rather than $Ax = b$. This won't cause much issue if A is well-conditioned, but could be a problem otherwise.

Perturbed linear system

Theorem

Suppose A is nonsingular and $\|\delta A\| < \frac{1}{\|A^{-1}\|}$, then the solution $\tilde{x}$ of perturbed linear system $(A + \delta A)x = b + \delta b$ has an error estimate given by

$$\frac{\|x - \tilde{x}\|}{\|x\|} \le \frac{K(A)\|A\|}{\|A\| - K(A)\|\delta A\|} \Big(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \Big)$$

where x is the solution of the original linear system $Ax = b$.

Note that $K(A)\|\delta A\| = \|A\|\|A^{-1}\|\|\delta A\| < \|A\|$ so the denominator is positive.

Conjugate gradient method

Conjugate gradient (CG) method is particularly efficient for solving linear systems with large, sparse, and positive definite matrix A.

Equipped with proper preconditioning, CG can often reach very good result in $\sqrt{n}$ iterations (n the size of system).

The per-iteration cost is also low when A is sparse.

An alternate perspective of linear system

Theorem

Let A be positive definite, then x^* is the solution of $Ax = b$ iff x^* is the minimizer of

$$g(x) = \frac{1}{2}x^\top Ax - b^\top x$$

Proof.

Note that $\nabla g(x) = Ax - b$ and $\nabla^2 g(x) = A \succ 0$, so $g(x^*) = Ax^* - b = 0$ iff x^* is a minimizer of $g(x)$. $\square$

An alternate perspective of linear system

We have following observations:

- $r = b - Ax = -\nabla g(x)$ is the residual and also the steepest descent direction of $g(x)$ (recall that $\nabla g(x)$ is the steepest ascent direction).
- It seems intuitive to update $x \leftarrow x + t \cdot r = x - t\nabla g(x)$ with proper step size t.
- It turns out that we can find such t that makes the most progress.
- This method is called the "steepest descent method".
- However, it converges slowly and exhibits "zigzag" path for ill-conditioned A.

A-orthogonal

Conjugate gradient method amends this issue of steepest descent. To derive CG, we first present the following concept:

Definition

*Two vectors v and w are called A-**orthogonal** if $\langle v, Aw \rangle = 0$.*

Theorem

If A is positive definite, then there exists a set of independent vectors $\{v^{(1)}, \cdots, v^{(n)}\}$ such that $\langle v^{(i)}, Av^{(j)} \rangle = 0$ for all $i \neq j$.

Key idea of CG

Given previous estimate $x^{(k-1)}$ and a "search direction" $v^{(k)}$, CG will find scalars t_k and s_k to update x and v:

$$
\begin{aligned}
x^{(k)} &= x^{(k-1)} + t_k v^{(k)} \\
v^{(k+1)} &= r^{(k)} + s_k v^{(k)}
\end{aligned}
$$

(where $r^{(k)} = b - Ax^{(k)}$), such that:

$$
\begin{aligned}
\langle v^{(k+1)}, Av^{(j)} \rangle &= 0, \quad \forall j \le k \\
\langle r^{(k)}, v^{(j)} \rangle &= 0, \quad \forall j \le k
\end{aligned}
$$

If this can be done, then $\{v^{(1)}, \cdots, v^{(n)}\}$ is A-orthogonal.

Derivation of t_k and s_k

The main tool is mathematical induction: given $x^{(0)}$, first set $v^{(0)} = 0$, $r^{(0)} = b - Ax^{(0)}$, $v^{(1)} = r^{(0)}$. So

$$\langle v^{(k+1)}, Av^{(j)} \rangle = 0, \quad \forall j \le k$$
$$\langle r^{(k)}, v^{(j)} \rangle = 0, \quad \forall j \le k$$

is true for $k = 0$. Assume they hold for $k - 1$, we need to find t_k and s_k such that they also hold for k.

Derivation of t_k and s_k

We first find t_k: note that

$$r^{(k)} = b - Ax^{(k)} = b - A(x^{(k-1)} + t_k v^{(k)}) = r^{(k-1)} - t_k A v^{(k)}$$

Therefore, by induction hypothesis, there is

$$\begin{aligned}
\langle r^{(k)}, v^{(j)} \rangle &= \langle r^{(k-1)} - t_k A v^{(k)}, v^{(j)} \rangle \\
&= \begin{cases} 0 & \text{if } j \le k-1, \\ \langle r^{(k-1)}, v^{(k)} \rangle - t_k \langle v^{(k)}, A v^{(k)} \rangle, & \text{if } j = k \end{cases}
\end{aligned}$$

So we just need

$$t_k = \frac{\langle r^{(k-1)}, v^{(k)} \rangle}{\langle v^{(k)}, A v^{(k)} \rangle}$$

to make $\langle r^{(k)}, v^{(j)} \rangle = 0$.

Derivation of t_k and s_k

Then we find s_k: by the update of $v^{(k+1)}$, we have

$$
\begin{aligned}
\langle v^{(k+1)}, Av^{(j)} \rangle &= \langle r^{(k)} + s_k v^{(k)}, Av^{(j)} \rangle \\
&= \begin{cases} \langle r^{(k)}, Av^{(j)} \rangle, & \text{if } j \le k-1 \\ \langle r^{(k)}, Av^{(k)} \rangle + s_k \langle v^{(k)}, Av^{(k)} \rangle, & \text{if } j = k \end{cases}
\end{aligned}
$$

Note that $Av^{(j)} = \frac{Ax^{(j)} - Ax^{(j-1)}}{t_j} = \frac{r^{(j-1)} - r^{(j)}}{t_j}$, and $r^{(j-1)} - r^{(j)}$ is linear combination of $v^{(j-1)}, v^{(j)}, v^{(j+1)}$, so $\langle r^{(k)}, Av^{(j)} \rangle = 0$ for $j \le k-1$ due to induction hypothesis. Hence we just need

$$
s_k = -\frac{\langle r^{(k)}, Av^{(k)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}
$$

to make $\langle r^{(k)}, Av^{(j)} \rangle = 0$ for all $j \le k$.

Derivation of t_k and s_k

We can further simply t_k and s_k:

Since that $v^{(k)} = r^{(k-1)} + s_{k-1}v^{(k-1)}$ and $\langle r^{(k-1)}, v^{(k-1)} \rangle = 0$, we have

$$t_k = \frac{\langle r^{(k-1)}, v^{(k)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} = \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle}$$

Since $r^{(k-1)} = v^{(k)} - s_{k-1}v^{(k-1)}$, we have $\langle r^{(k)}, r^{(k-1)} \rangle = 0$. Since $Av^{(k)} = \frac{Ax^{(k)} - Ax^{(k-1)}}{t_k} = \frac{r^{(k-1)} - r^{(k)}}{t_k}$, we have $\langle r^{(k)}, Av^{(k)} \rangle = -\frac{\langle r^{(k)}, r^{(k)} \rangle}{t_k}$. Combining t_k expression above, we have

$$s_k = -\frac{\langle r^{(k)}, Av^{(k)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} = -\frac{-\frac{\langle r^{(k)}, r^{(k)} \rangle}{t_k}}{\frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{t_k}} = \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle r^{(k-1)}, r^{(k-1)} \rangle}$$

Conjugate gradient method

Since $\langle r^{(n)}, v^{(k)} \rangle = 0$ for all $k = 1, \dots, n$ and the A-orthogonal set $\{v^{(1)}, \cdots, v^{(n)}\}$ is independent when A is positive definite, we know $r^{(n)} = b - Ax^{(n)} = 0$, i.e., $x^{(n)}$ is the solution.

This shows that CG converges in at most n steps, assuming all arithmetics are exact.

Conjugate gradient method

- Input: $x^{(0)}$, $r^{(0)} = b - Ax^{(0)}$, $v^{(1)} = r^{(0)}$.
- Repeat the following for $k = 1, \dots, n$ until $r^{(k)} = 0$:

$$\begin{aligned}
t_k &= \frac{\langle r^{(k-1)}, r^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} \\
x^{(k)} &= x^{(k-1)} + t_k v^{(k)} \\
r^{(k)} &= r^{(k-1)} - t_k A v^{(k)} \\
s_k &= \frac{\langle r^{(k)}, r^{(k)} \rangle}{\langle r^{(k-1)}, r^{(k-1)} \rangle} \\
v^{(k+1)} &= r^{(k)} + s_k v^{(k)}
\end{aligned}$$

- Output: $x^{(k)}$.

Preconditioning

The convergence rate of CG can be greatly improved by **preconditioning**. Preconditioning reduces condition number of A first if A is ill-conditioned. With preconditioning, CG usually converges in $\sqrt{n}$ steps.

The preconditioning is done by using some nonsingular matrix C, we can get $\tilde{A} = C^{-1}A(C^{-1})^\top$ such that $K(\tilde{A}) \ll K(A)$.

Now by defining $\tilde{x} = C^\top x$ and $\tilde{b} = C^{-1}b$, we obtain a new linear system $\tilde{A}\tilde{x} = \tilde{b}$, which is equivalent to $Ax = b$. Then we can apply CG to the new system $\tilde{A}\tilde{x} = \tilde{b}$.

Preconditioner

There are various methods to choose the preconditioner C.

- Choose $C = \text{diag}(\sqrt{a_{11}}, \dots, \sqrt{a_{nn}})$.
- Approximate Cholesky's factorization $LL^\top \approx A$ (by ignoring small values in A) and set $C = L$ (then $C^{-1}A(C^{-1})^\top \approx L^{-1}(LL^\top)L^{-T} = I$).
- Many others...

Preconditioned conjugate gradient method

- Input: Preconditioner C, $x^{(0)}$, $r^{(0)} = b - Ax^{(0)}$, $w^{(0)} = C^{-1}r^{(0)}$, $v^{(1)} = C^{-T}w^{(0)}$.
- Repeat the following for $k = 1, \dots, n$ until $r^{(k)} = 0$:

$$\begin{aligned}
\tilde{t}_k &= \frac{\langle w^{(k-1)}, w^{(k-1)} \rangle}{\langle v^{(k)}, Av^{(k)} \rangle} \\
x^{(k)} &= x^{(k-1)} + \tilde{t}_k v^{(k)} \\
r^{(k)} &= r^{(k-1)} - \tilde{t}_k A v^{(k)} \\
w^{(k)} &= C^{-1} r^{(k)} \\
\tilde{s}_k &= \frac{\langle w^{(k)}, w^{(k)} \rangle}{\langle w^{(k-1)}, w^{(k-1)} \rangle} \\
v^{(k+1)} &= C^{-\top} w^{(k)} + \tilde{s}_k v^{(k)}
\end{aligned}$$

- Output: $x^{(k)}$.

A comparison

Example

Given A and b below, we use the methods above to solve $Ax = b$.

$$A = \begin{bmatrix} 0.2 & 0.1 & 1 & 1 & 0 \\ 0.1 & 4 & -1 & 1 & -1 \\ 1 & -1 & 60 & 0 & -2 \\ 1 & 1 & 0 & 8 & 4 \\ 0 & -1 & -2 & 4 & 700 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

True solution is

$$x^* = \begin{bmatrix} 7.859713071 \\ 0.4229264082 \\ -0.07359223906 \\ -0.5406430164 \\ 0.01062616286 \end{bmatrix}$$

A comparison

A comparison of Jacobi, Gauss-Seidel, SOR, CG, and PCG on the problem above.

Method	Number of Iterations	$\mathbf{x}^{(k)}$	$\lVert \mathbf{x}^* - \mathbf{x}^{(k)} \rVert_\infty$
Jacobi	49	$(7.86277141, 0.42320802, -0.07348669, -0.53975964, 0.01062847)^t$	0.00305834
Gauss-Seidel	15	$(7.83525748, 0.42257868, -0.07319124, -0.53753055, 0.01060903)^t$	0.02445559
SOR ($\omega = 1.25$)	7	$(7.85152706, 0.42277371, -0.07348303, -0.53978369, 0.01062286)^t$	0.00818607
Conjugate Gradient	5	$(7.85341523, 0.42298677, -0.07347963, -0.53987920, 0.008628916)^t$	0.00629785
Conjugate Gradient (Preconditioned)	4	$(7.85968827, 0.42288329, -0.07359878, -0.54063200, 0.01064344)^t$	0.00009312